Winning Space Race with Data Science

Michael Lu
March 29, 2022

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion
- Appendix

Executive Summary

Methodology

Space X launch data was gathered directly from the Space X API and the Space X Wikipedia page. Data was cleaned and prepared for exploratory data analysis in search of relationships and patterns. Exploratory data analysis was done using statistics, math, data visualization (graphs, charts, maps), and by building an interactive web application. Relevant data was identified and categorized, and the data set was used to train machine learning models in order to perform predictive analysis and identify characteristics which have a large impact on first stage landing success.

Results

Initial data analysis reveals that Space X has gradually increased its competitive advantage in the market by successfully landing and reusing the first stage even more frequently. Four machine learning models were trained on the data sets, and all produced similar results on a relatively small sample size (18). While the initial results are promising, more data is needed.

Introduction

Background

- The era of commercial space exploration has dawned upon us. Space X is the most successful venture thus far, in large part due to its relative affordability.
- Space X offers *Falcon 9* rocket launches at a cost of $69 million (USD) -- a distinct competitive edge over the competition, which costs upwards of $165 million (USD)
- Space X's relative affordability is due in large part to its first stage's reusability.
- First stage reusability is a "must", if any company wants to stand a chance of competitively penetrating the market.
- Using data science, we can determine if the first stage will land, subsequently allowing the company to determine the total cost of launch.

Questions

- Which rocket characteristics have the most impact on successful landing rate?
- Which conditions (geographical or otherwise) give the best chance of a successful landing?

Section 1

Methodology

Methodology

- **Data Collection** – sourced Space X mission data directly from the API and through web scraping the Wikipedia page
- **Data Wrangling** – refined/cleaned the data set in preparation for exploratory data analysis
- **Exploratory Data Analysis (EDA)** – analyzed data using mathematical/statistical queries and leveraging visualization libraries to begin uncovering relationships/patterns in the data set
- **Interactive Visual Analytics** – designed an interactive web-based application to provide a more dynamic visual presentation of the results using malleable parameters to display data visualizations of select launches
- **Predictive Analysis** – used machine learning to create a predictive model based on the data set to predict likely outcomes of future launches

Data Collection

- Data was collected both directly from the Space X API and by web scraping data from the Space X Wikipedia page.
  - **API** – Application Programming Interface is a type of software interface that provides services to other software. It connects computers or pieces of software to each other. In our case, it connects us to the Space X database. Essentially, the Space X API functions as a memory bank for Space X launch history. We call on the Space X API to request data from the Space X database. The API returns the data for analysis.
  - **Web Scraping** is the process of automatically collecting/mining data from a website. For our applications, we use the technique to rip the data from the table in the Wikipedia page, "List of Falcon 9 and Falcon Heavy Launches"

Data Collection - Scraping

- Request the HTML data from the Wikipedia page: "List of Falcon 9 and Falcon Heavy Launches".
- Use the web scraper, BeautifulSoup to find and extract the tables from the HTML as a dictionary, before converting the dictionary to a data frame for analysis.

GitHub

Data Wrangling

- Falcon 1 launches were removed from the data set to focus on Falcon 9 launches.
- In order to more easily perform meaningful data analysis, categorical variables were converted to numerical values, or "dummies".
- For our purposes, primarily interested in the success rate of landing.
- For landing outcomes, we assigned "success" = 1 and "fail" = 0
  - Results: "True ASDS", "True RTLS", "True Ocean" = 1
  - Results: "None None", "False ASDS", "None ASDS", "False Ocean", "False RTLS" = 0

GitHub

EDA with Data Visualization

- Scatter Plots are an excellent visualization of how one variable affects another.
  - *Flight Number vs. Launch Site*
  - *Flight Number vs. Payload Mass*
  - *Payload Mass vs. Launch Site*
  - *Payload Mass vs. Orbit Type*
- Bar Charts allow easy comparison across categories.
  - *Success Rate vs. Orbit Type*
- Line Charts show variable correlations and facilitate predictions of future instances.
  - *Success Rate vs. Year*

GitHub

EDA with SQL

- Loaded wrangled/cleaned data from both the Space X API and Space X Wiki page into the IBM Db2 Database.
- Queried the data for analysis using SQL Python integration to communicate with the database.
  - Explored data, querying information about launch site names, first instances, mission outcomes, pay load sizes, customers, and booster versions.
  - Generated lists, rankings, and counts of the above mentioned variables for analysis.

GitHub

Build an Interactive Map with Folium

- Created interactive maps using Folium, marking launch sites and including data such as:
  - The number of successful and unsuccessful landings from the launch sites
  - Proximity of launch sites to key locations, including railways, highways, coasts, and cities
- These maps revealed trends in where launch sites were located, and whether the location of each launch site specifically impacts landing success rate.

GitHub

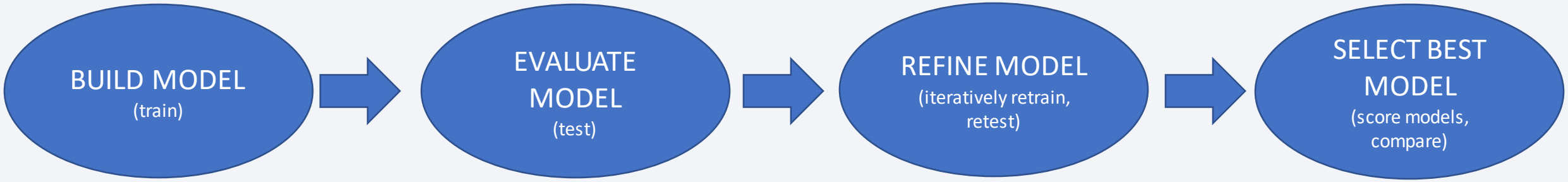
Build a Dashboard with Plotly Dash

- Built a dashboard, an interactive web application, to visually compare data across launch/landing instances.
- The dashboard composed of two visualizations:
  - PIE CHART: shows distribution of successful landings across launch sites
  - SCATTER PLOT: shows relationship between success rate across launch sites depending on payload mass, based on two inputs:
    - All Sites/Individual Site
    - Payload Mass – sliding scale 0 – 10,000kg

GitHub

Predictive Analysis (Classification)

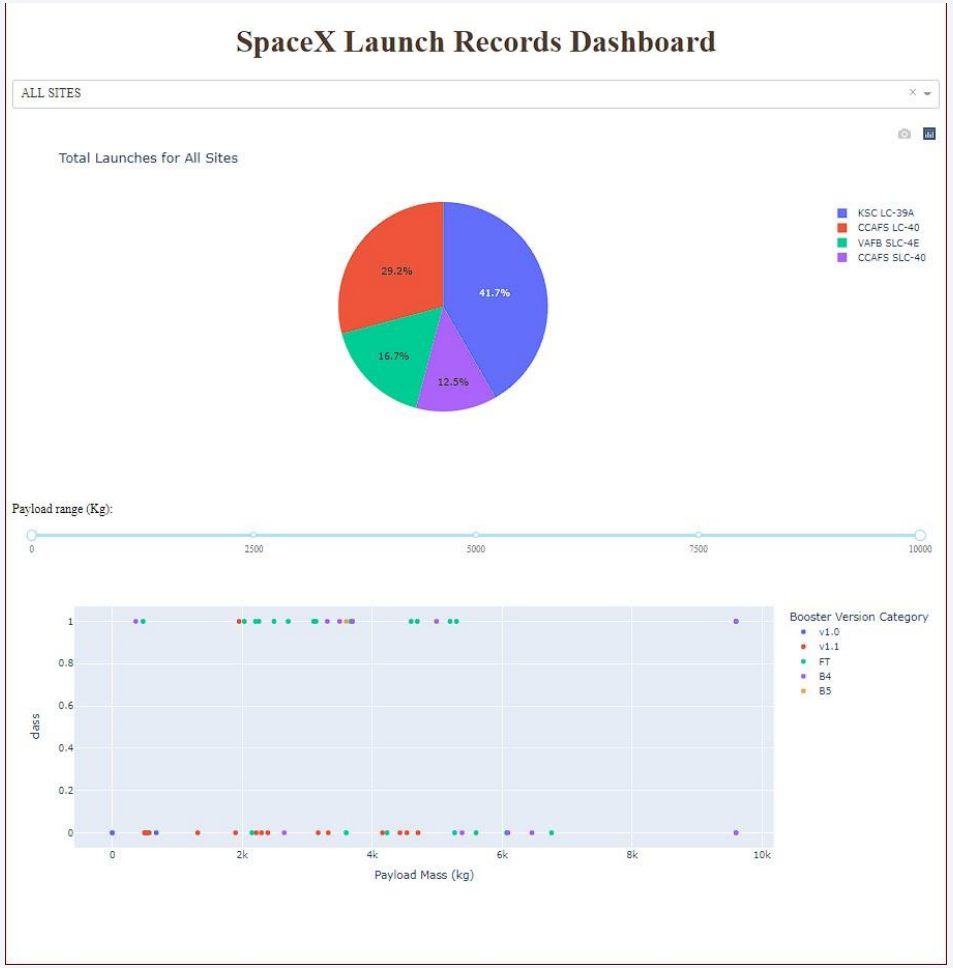
- Using the results from EDA, determined what data to use to train the machine learning algorithms for further analysis.
  - Created column class and standardized data
  - Split data into training and testing
  - Iteratively trained and tested each model (Support Vector Machine (SVM), Classification Tree, Logistic Regression Model) to minimize model error (find accurate hyperparameters)



GitHub

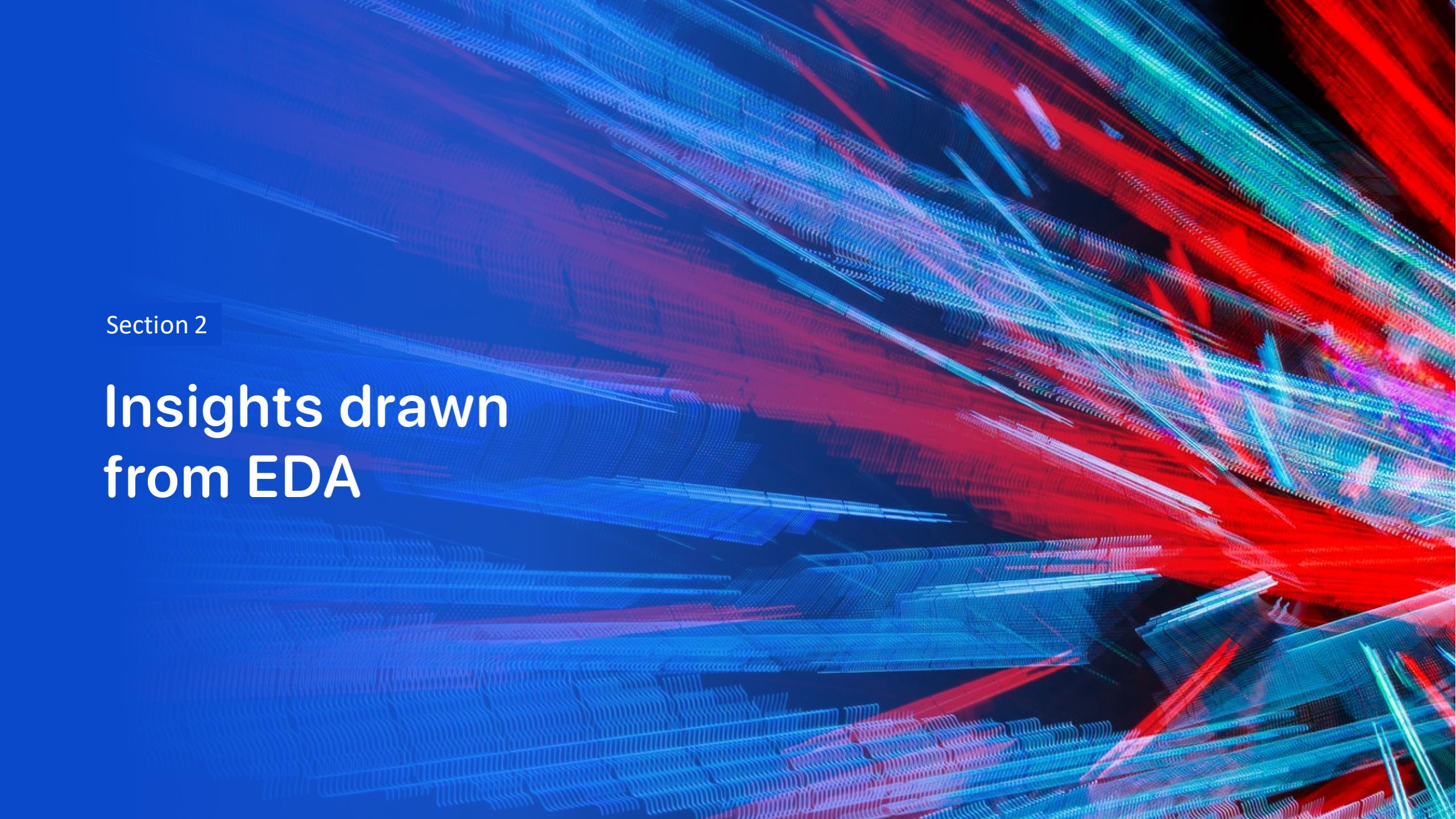
Results

- EDA revealed the presence or absence of correlation between the above-mentioned variables, detailed in the following slides
  - Shown here, a screenshot of the interactive dashboard used for EDA
- Predictive Analysis showed that each of the models was equally accurate (~83%) at predicting outcomes.
  - It should be noted that the sample size (18) is small, and more data is necessary.



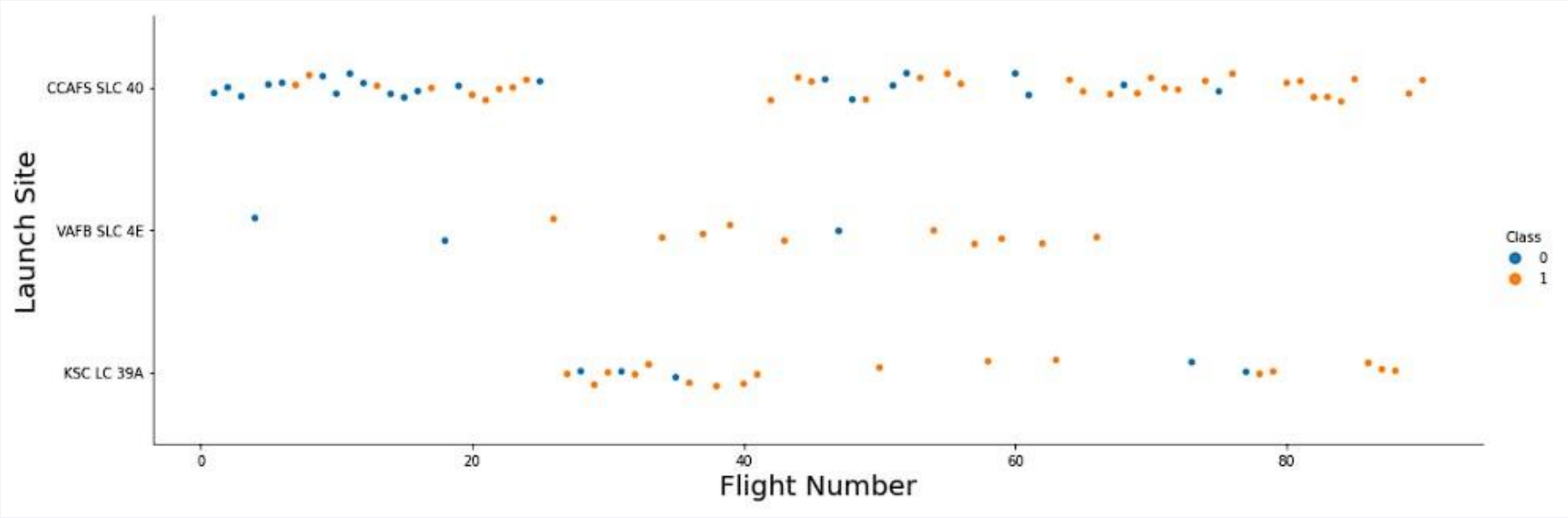
Section 2

Insights drawn from EDA



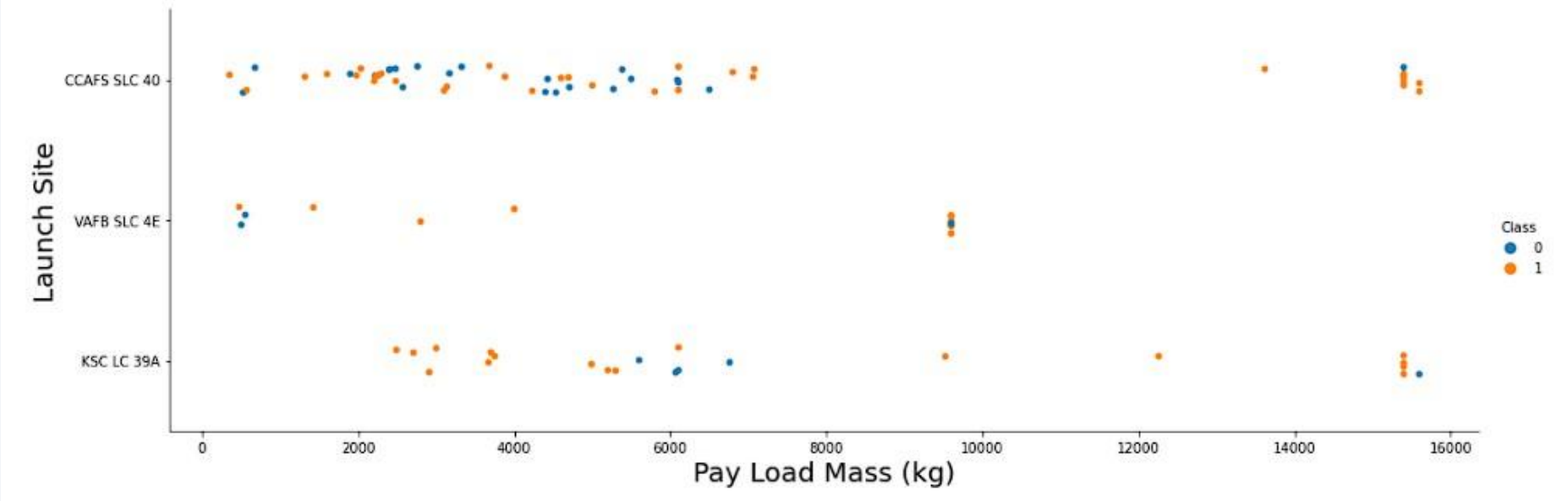
Flight Number vs. Launch Site

- CLASS 0 - failure
- CLASS 1 – success
- Figure reveals that success rate has increased as time has passed (more flights)
  - Flight #20 is of note, as flight success rate seems to increase from this point on



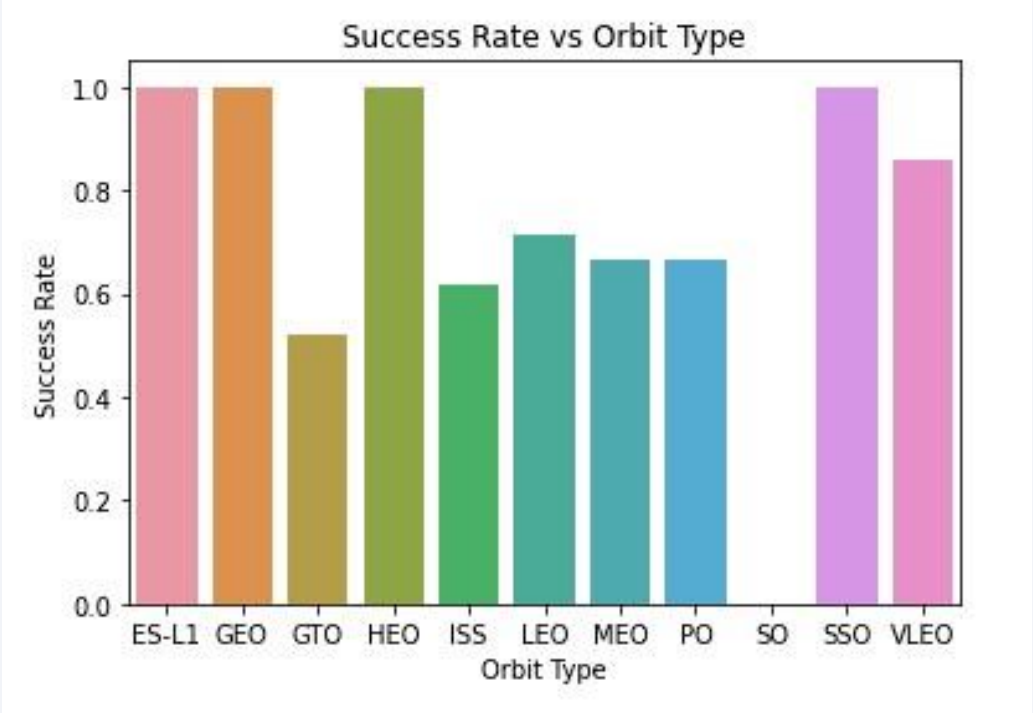
Payload vs. Launch Site

- There seems to be a weak trend of higher success at higher payload mass
- Generally, payload mass ranges under 8,000kg.
- There is no observable pattern in the data set, and no conclusions can be drawn.



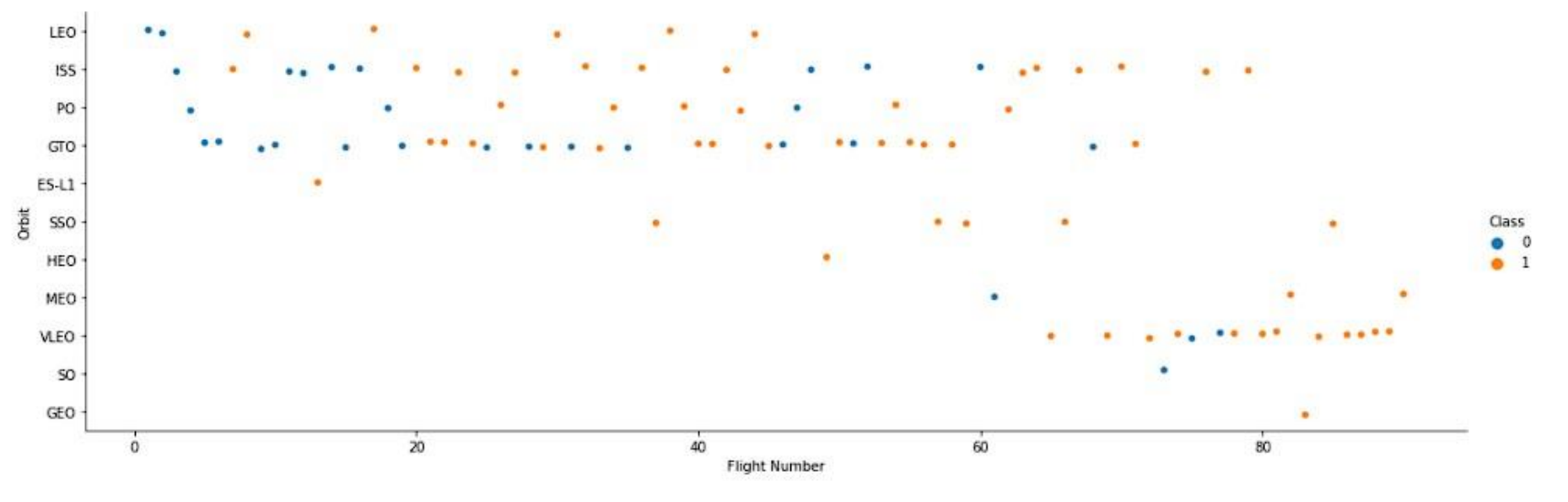
Success Rate vs. Orbit Type

- Orbit Types ES-L1, GEO, HEO, and SSO have the highest success rate (100%) but very small sample size
- Orbit Type SO has a success rate of 0%
- Note that launch types SO, GEO, HEO, ES-L1 only have 1 instance, each
- The success rate of GTO is ~50%, with the largest sample size of 27 instances
  - ISS at ~60% (21 instances)
  - VLEO at ~85% (14 instances)



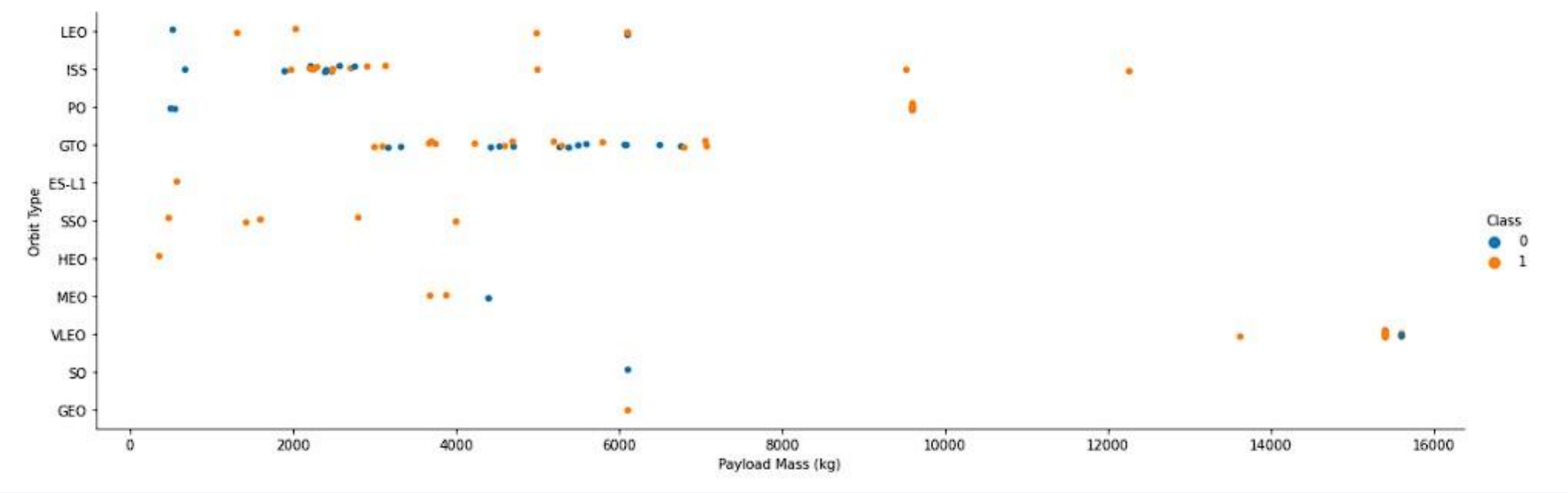
Flight Number vs. Orbit Type

- Generally, preference of orbit type changed over time
- Success rate has increased over time
- This correlation is likely indicative of favoring more successful orbit types
- Generally, all orbit types' success rates have increased over time



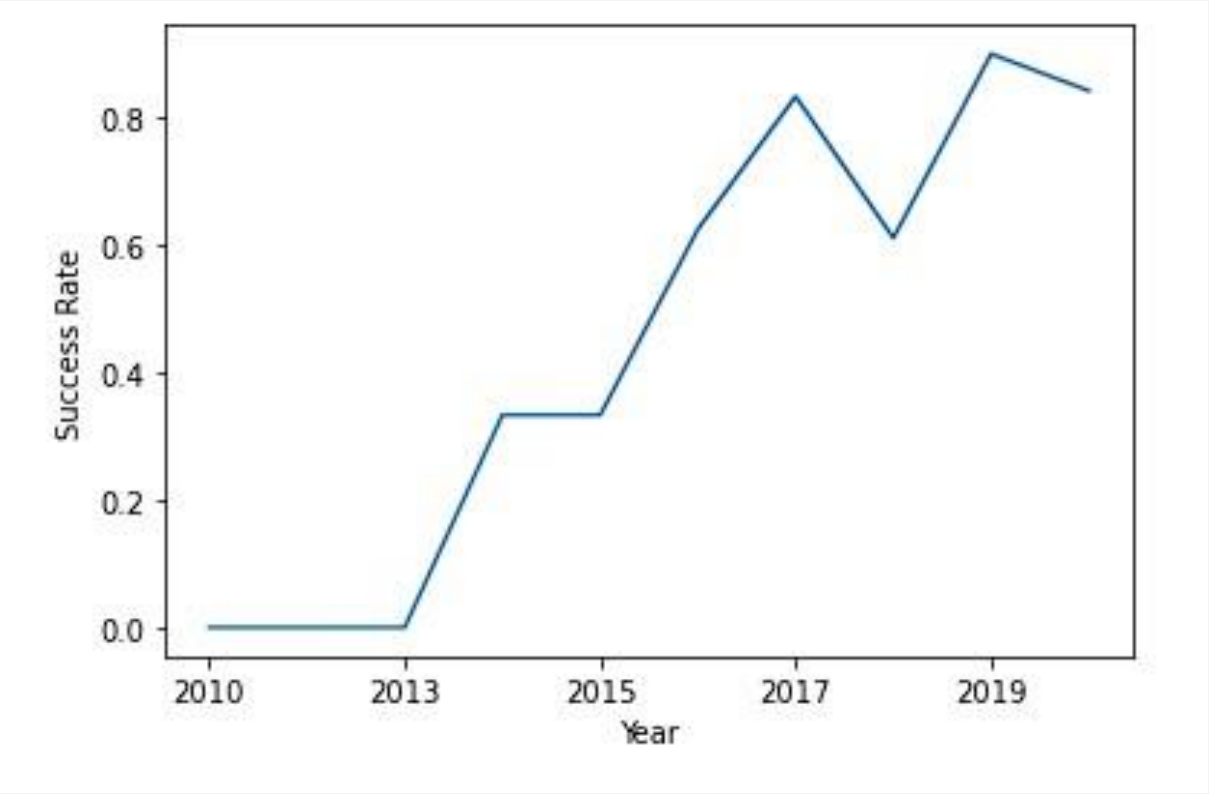
Payload vs. Orbit Type

- Heavier payloads generally have higher success rate
- Heavier payloads (9k–13k kg) take ISS and PO orbit types with 100% success
- Heaviest payloads (13k—16k kg) take VLEO orbit type with high success



Launch Success Yearly Trend

- Success rate has generally increased since 2013
- There was a slight decrease in success rate in 2018
- Most recent success rate hovering at/above 80% (2019-2020)



All Launch Site Names

- Using the SQL "DISTINCT" clause on the column "LAUNCH_SITE" displays only unique values in the generated list

```
%%sql
select distinct LAUNCH_SITE from SPACEXTBL
```

Done.

launch_site
CCAFS LC-40
CCAFS SLC-40
KSC LC-39A
VAFB SLC-4E

Launch Site Names Begin with 'CCA'

- Called all records where the launch site started with "CCA" by using the "like" operator combined with %
- Limited record display to first 5 entries

```
%sql select * from SPACEXTBL where LAUNCH_SITE like 'CCA%' \
limit 5
```

[redacted]

Done.

DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- Used the "sum()" function to add up the total payload mass where the record's customer was listed as NASA
- NASA sent a total of 45,596 kg of payload to space using the Space X Falcon 9

```
%sql select sum(PAYLOAD_MASS__KG_) as total_payload_mass_KG from SPACEXTBL \
where CUSTOMER = 'NASA (CRS)'
```

Done.

total_payload_mass_kg
45596

Average Payload Mass by F9 v1.1

- Used the "avg()" function to calculate the average payload mass where the record's booster version was listed as the "F9 v1.1"
- The average payload carried by the F9 v1.1 booster was 2928 kg

```
%sql select avg(PAYLOAD_MASS__KG_) as average_payload_mass_f9_v1_1_kg from SPACEXTBL \
where BOOSTER_VERSION = 'F9 v1.1'
```

Done.

average_payload_mass_f9_v1_1_kg
2928

First Successful Ground Landing Date

- Used the "min()" function to find the minimum/earliest date where the record's landing outcome was listed as successful on a ground pad
- The first successful ground landing was on December 22, 2015

```
%sql select min(DATE) as first_ground_pad_success_landing_date from SPACEXTBL \
where LANDING__OUTCOME = 'Success (ground pad)'
```

Done.

first_ground_pad_success_landing_date
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- Called the booster versions which carried a payload between 4000-6000 kg which successfully landed on a drone ship
- All boosters were of the F9 FT series:
  - B1022
  - B1026
  - B1021.2
  - B1031.2

```
%sql select BOOSTER_VERSION from SPACEXTBL \
where PAYLOAD_MASS__KG_ between 4000 and 6000 \
and LANDING__OUTCOME='Success (drone ship)'
```

Done.

booster_version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

- Used the "count()" function and "group by" statement to sum up the total number of successful and failed mission outcomes
- Space X has an over 99% success rate

```
%sql select MISSION_OUTCOME, count(*) as total_count_outcome from SPACEXTBL \
group by MISSION_OUTCOME
```

Done.

mission_outcome	total_count_outcome
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- Calls a list of the booster versions which have carried the maximum payload mass
- Max payload mass is 15,600kg
- 12 different Falcon 9 boosters have successfully carried the maximum payload

```
%sql select BOOSTER_VERSION from SPACEXTBL \
where PAYLOAD_MASS__KG_ = \
(select max(PAYLOAD_MASS__KG_) from SPACEXTBL)
```

Done.

booster_version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

2015 Launch Records

- Calls a table listing landing outcome, booster version, and launch site from 2015 where the landing attempt failed
- Only 2 landing attempts failed in 2015, both from the same launch site

```
%sql select LANDING__OUTCOME, BOOSTER_VERSION, LAUNCH_SITE from SPACEXTBL \
where LANDING__OUTCOME = 'Failure (drone ship)' \
and YEAR(DATE) = 2015
```

Done.

landing__outcome	booster_version	launch_site
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

- Called a table which listed landing outcome and the number of occurrences between the dates of June 4, 2010 and March 20, 2017.
- Ordered the list in descending order.
- Most often, no attempt was made at landing the first stage.
- When landing was attempted, success and failure rates were similar.

```
%sql select LANDING__OUTCOME, COUNT(LANDING__OUTCOME) as COUNT from SPACEXTBL \
where DATE between '2010-06-04' and '2017-03-20' \
group by LANDING__OUTCOME \
order by COUNT desc
```

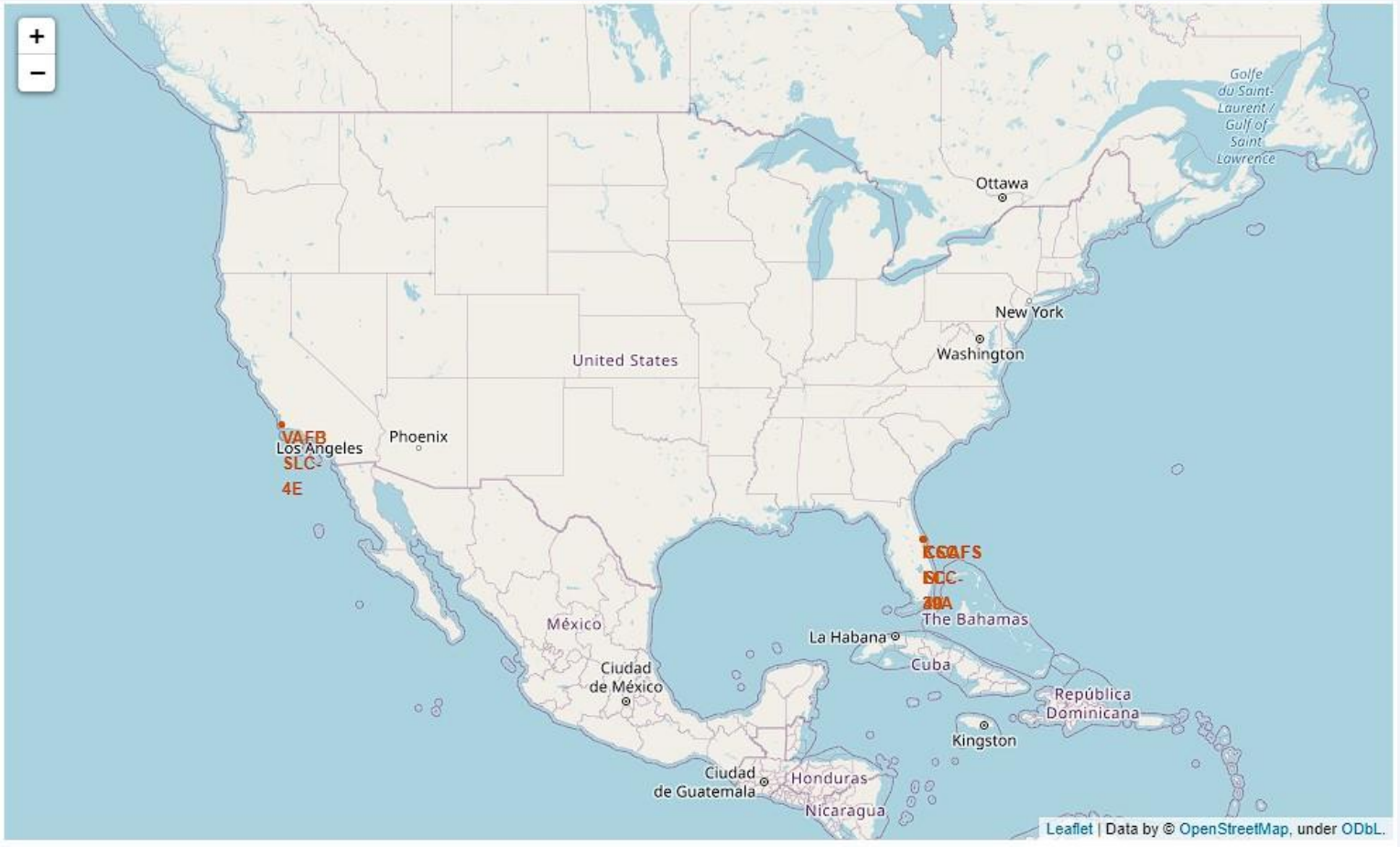
Done.

landing__outcome	COUNT
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1


Section 3
Launch Sites
Proximities Analysis

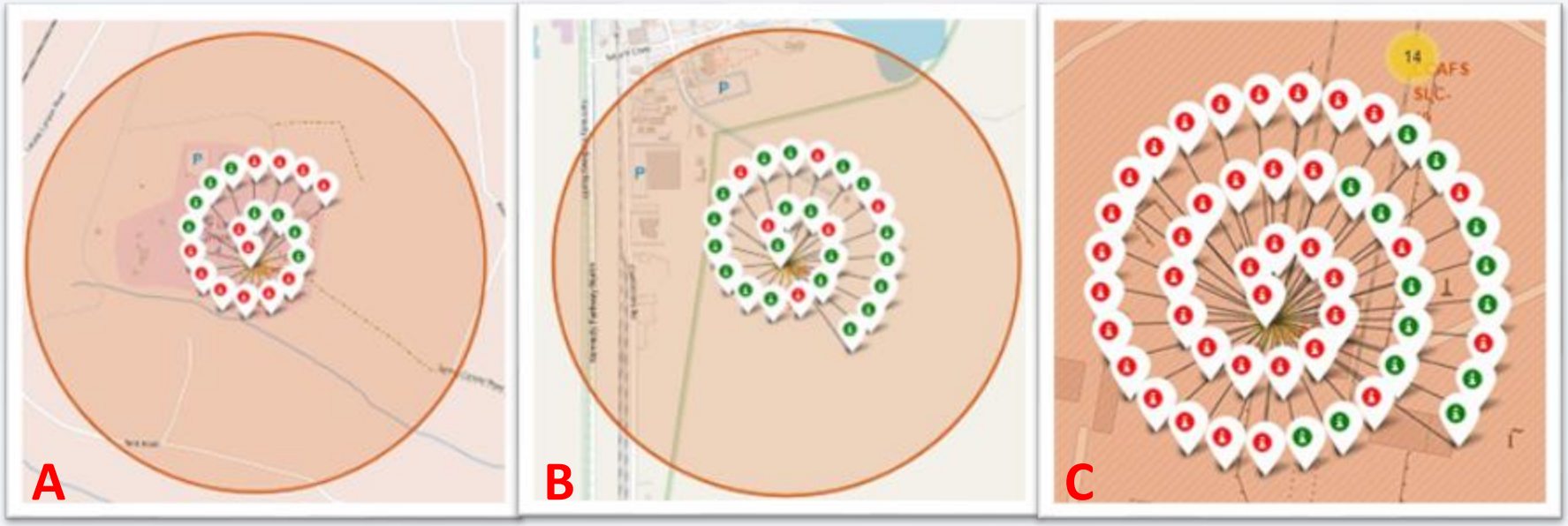
Space X Launch Site Locations

- All launch sites in USA
- All launch sites near ocean, in southern half of USA



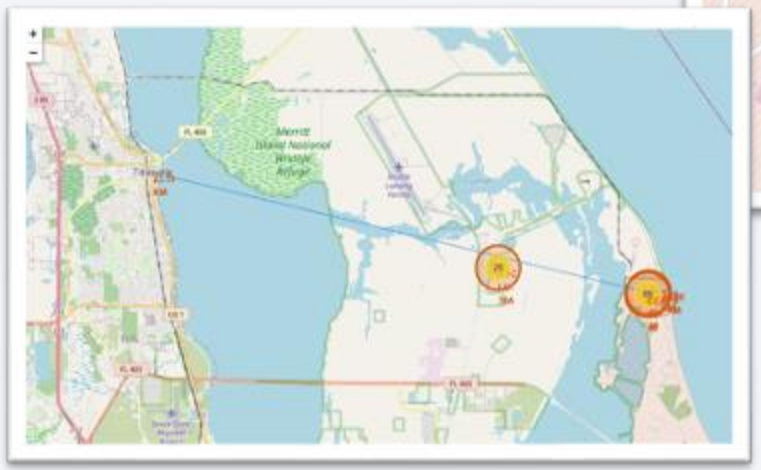
Color-coded Clusters of Launch Markers

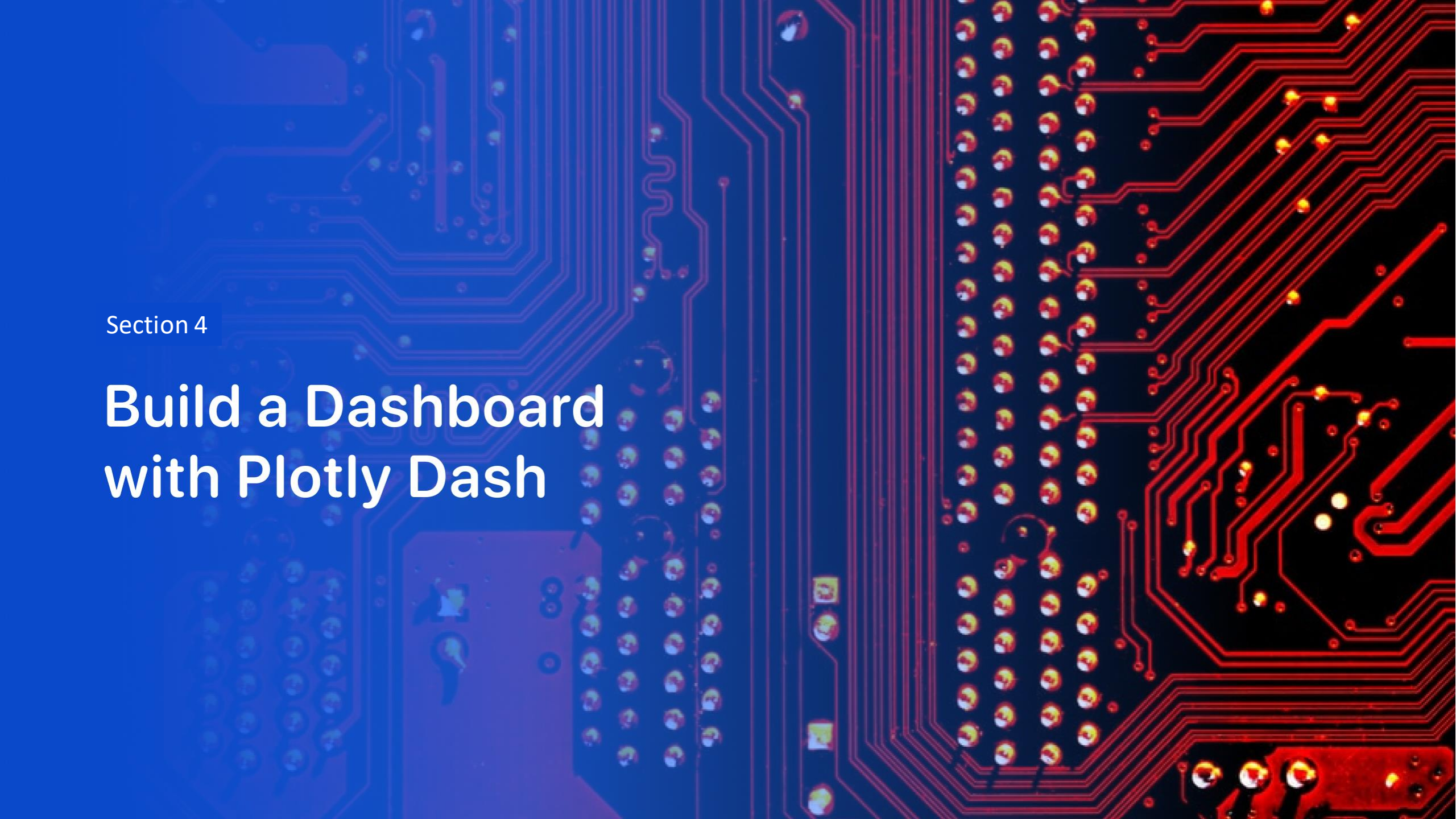
- Zooming in and clicking on launch sites drills down and displays successful and failed launches at each site:
  - A: VAFB SLC-4e – Vandenberg AFB Space Launch Complex 4e – California, USA
  - B: KSC LC-39a – Kennedy Space Center Launch Complex 39a – Florida, USA
  - C: CCAFS (S)LC-40 – Cape Canaveral Space Launch Complex 40 – Florida, USA



Launch Site Proximity to Key Locations

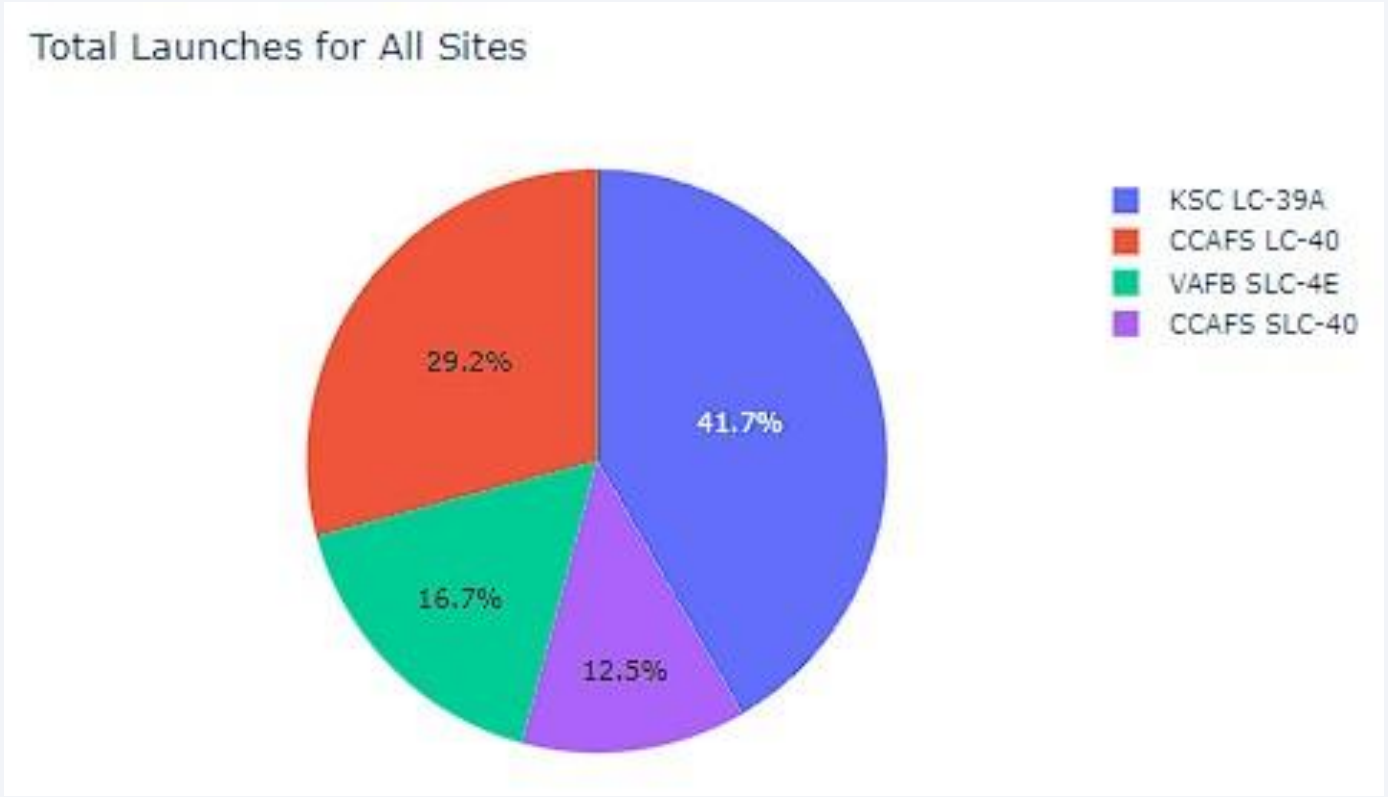
- Using CCAFS SLC-40 as an example, most launch sites tend to be very close to highways, railways, and the ocean, while being further away from cities.
- Proximity to highways and railways allows for convenient transport of supplies.
- Proximity to the ocean and distance from cities ensures that, failed launches land in the ocean and far away from human population centers.




Section 4
Build a Dashboard with Plotly Dash

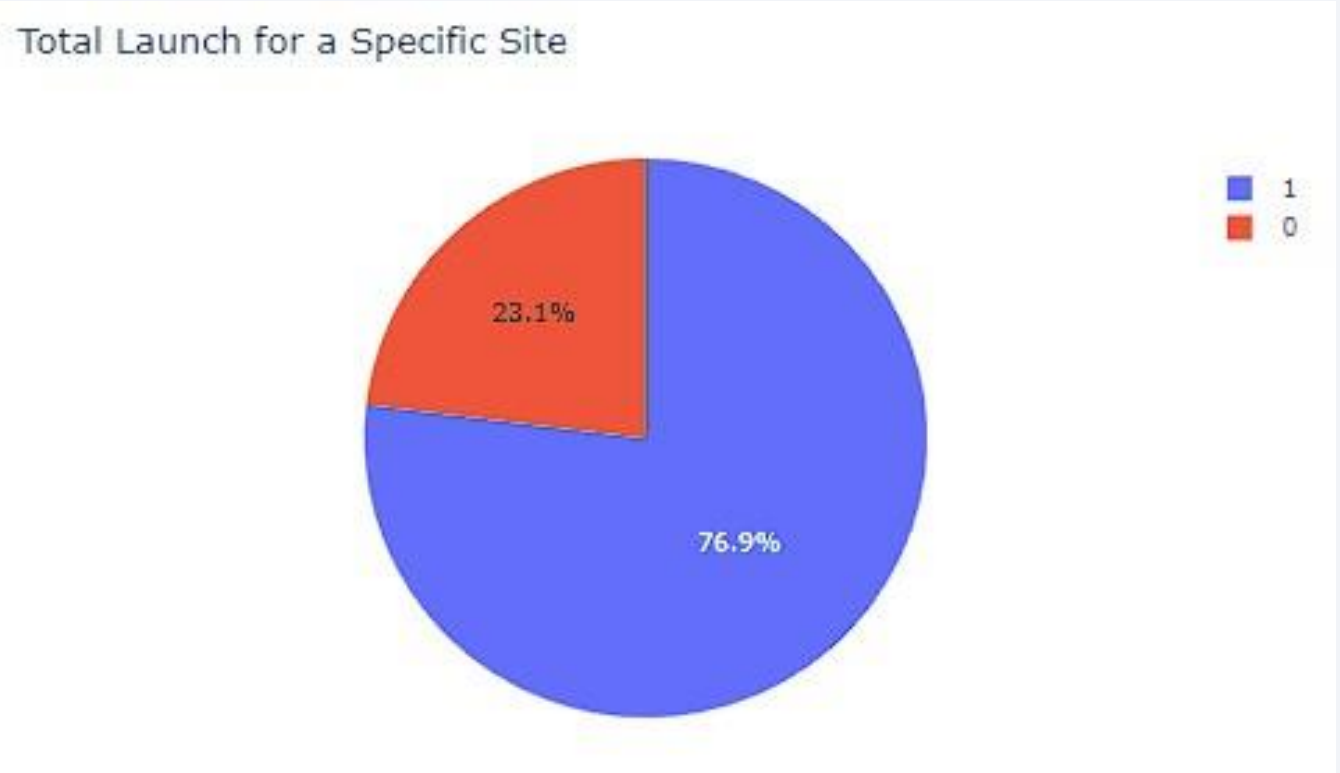
Distribution of Successful Launches – All Sites

- CCAFS records both the highest volume of launches and successes (45.9%), followed by:
  - KSC (41.7%)
  - VAFB 16.7%
- Note that VAFB is the only launch site on the west coast and experiences the lowest volume of launches



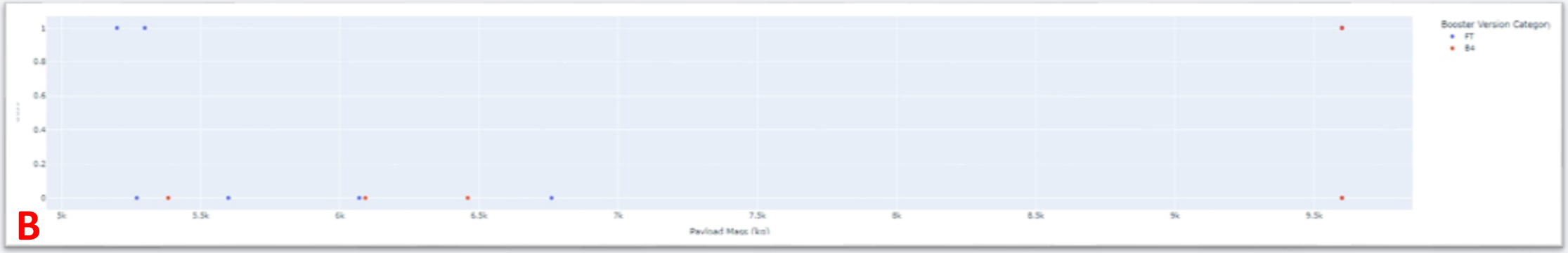
Most Successful Launch Site – KSC LC-39a

- KSC LC-39a has the highest rate of successful launches at 76.9% (10 successes), versus 23.1% (3) failures



Launch Outcome vs. Payload Mass by Booster

- The interactive visualization dashboard allows for clearer visualization of success rate versus payload mass by using a sliding scale to limit results to certain mass ranges
- Below are the launch outcomes versus payload weight for two ranges:
  - A: 0-5000kg
  - B: 5000-10,000kg

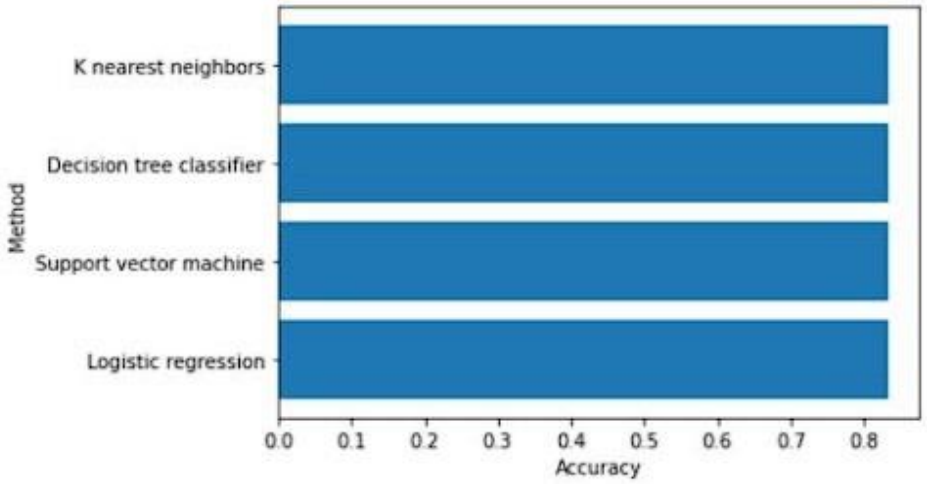


Section 5

Predictive Analysis (Classification)

Classification Accuracy

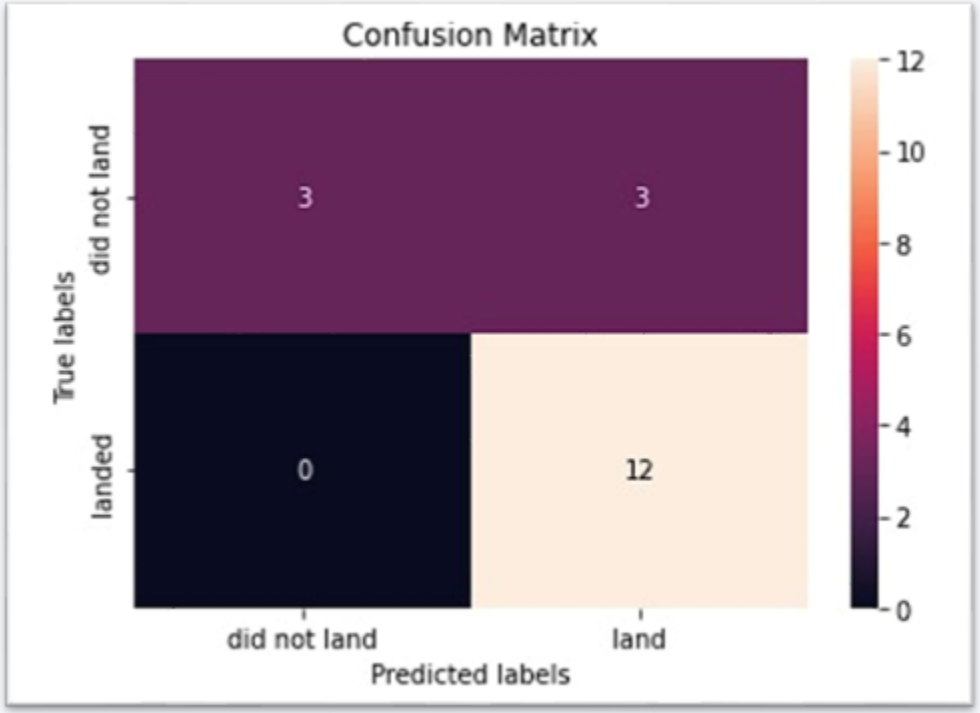
- All 4 prediction models have a model accuracy score of 83.3%, making them effectively identical
- While the model accuracy is encouraging, it must be noted that sample size was 18
  - It is likely that the small sample size is creating variance in the error results
  - More data is likely needed to create and test more accurate models



	Method	Accuracy
0	Logistic regression	0.833333
1	Support vector machine	0.833333
2	Decision tree classifier	0.833333
3	K nearest neighbors	0.833333

Confusion Matrix

- Since classification accuracy was identical, it is unsurprising that the confusion matrices for all 4 models are identical
- The models accurately predicted 12 successful and 3 failed landings
- The models mistakenly predicted 3 successful landings which actually failed
  - *False positives*
- The models overwhelmingly predict successful landings



Conclusions

- Space X launches have become increasingly successful
- Space X has gotten better at successfully launching heavier payloads, although generally, the success rate is higher in lighter payloads
- Space X is more frequently selecting orbital types with a high success rates
- Most launches occur in Florida, and an increased proportion of them are successful in comparison to launches in California
- Launch sites are near highways, railways, and the ocean – but far from cities
- ML models predict mostly successful launches, with high (83.3%) accuracy across all models – albeit with a small (18) sample size

Appendix

- GitHub
- Course Website: Applied Data Science Capstone
- Instructors: IBM Data Science Professional Certificate

Thank you!